# Artificial intelligence in human resources management: Challenges and a path forward

Peter Cappelli,[1] Prasanna Tambe,[1] and Valery Yakubovich[2]

## Abstract

We consider the gap between the promise and reality of AI in human resource management and suggest how progress might be made. We identify four challenges in using data science techniques in HR practices: 1) complexity of HR phenomena, 2) constraints imposed by small data sets, 3) ethical questions associated with fairness and legal constraints, and 4) employee reaction to AI-management. We propose practical responses to these challenges and converge on three overlapping principles--causal reasoning, randomization, and process formalization—that can potentially lead to both economically efficient and socially appropriate AI-management of human resources.

## Introduction

The speed with which the rhetoric about a digital transformation in management moved from big data (BD) to machine learning (ML) to artificial intelligence (AI) is staggering. The match between the rhetoric and reality is a different matter, however. Most companies are struggling to make any progress building data analytics capabilities: 41% percent of CEOs report that they are not at all prepared to make use of new data analytic tools, and only 4 percent say that they are "to a large extent" prepared (IBM 2018). There have been major advances in the domains of pattern recognition and natural language processing (NLP) over the last several years, and deep learning using neural networks, which has become increasingly common in some data-rich contexts, has brought us closer to true AI, which represents the ability of machines to mimic adaptive human decision making. Nevertheless, with respect to the management of employees, where the promise of more sophisticated decisions has been articulated loudly and often, few organizations have even entered the big data stage. Only 22 percent of firms say they have adopted analytics in human resources (LinkedIn 2018).

[1] The Wharton School, University of Pennsylvania
[2] ESSEC Business School, France

The promise of data analytics is arguably easiest to see in fields like marketing. While there are many questions to be answered there, they tend to be distinguished by their relative clarity, such as, what predicts who will buy a product or how changes in its presentation affect its sales. Outcomes are easily measured, are often are already collected electronically by the sales process, and the number of observations – sales of a particular item across the country over time, e.g. – is very large, making the application of big data techniques feasible. Although marketing is not without its ethical conundrums, the idea that companies should be trying to sell more of their products is well-accepted as is the idea that business will attempt to influence customers to buy more.

The effective application of AI to human resources problems presents very different challenges, despite the fact that virtually all of the efforts to improve employee management make use of data analytics, specifically algorithms to predict outcomes, such as who will be a good hire.

- A first problem is complexity of the outcomes of HR, such as what constitutes being a "good employee." There are many dimensions to that construct, and measuring it with precision for most jobs is quite difficult: performance appraisal scores, the most widely-used metric, have been roundly criticized for problems of validity and reliability as well as for bias, and many employers are giving them up altogether (Cappelli and Tavis 2017). Any reasonably complex job is interdependent with other jobs and therefore individual performance is hard to disentangle from group performance (Pfeffer and Sutton 2006).

- Second, the data sets in human resources tend to be quite small by the standards of data science. The number of employees that even a large company may have is trivial compared to the number of purchases their customers make, for example. Moreover, many outcomes of interest are rarely observed, such as who is fired for poor performance. Data science techniques perform poorly when predicting relatively rare outcomes.

- Third and arguably most important, the outcomes of human resource decisions, such as who gets hired and fired, have such serious consequences for individuals and society that elaborate legal frameworks are required to govern how employers must go about making those decisions, raising issues of procedural justice. Society is also concerned about the outcomes per se, raising issues of distributive justice. Employment decisions are also subject to a range of complex socio-psychological concerns, such as personal worth and status, perceived fairness, contractual and relational expectations, that affect organizational outcomes as well as individual ones. As a result, being able to explain and also to justify the practices one uses is much more important than in other fields.
- Finally, employee reactions may be a concern, as workers are active participants in a company's operations, capable of gaming or adversely reacting to algorithmic-based decisions.

To illustrate these concerns, consider the use of an algorithm to predict who to hire based on the attributes of good employees in the current workforce. Even if we could demonstrate a causal relationship between sex and job performance, we might well not trust an algorithm that says hire more white men because job performance itself may be a biased indicator, the attributes of the current workforce may be distorted by how we hired in the past (e.g., we hired few women), and both the legal system and social norms would create substantial problems for us if we did act on it.  If we instead build an algorithm on a more objective measure, such as who gets dismissed for poor performance, the number of such cases in a typical company is too small to construct an effective algorithm. Moreover, once applicants discover the content of our hiring algorithm (e.g., that it uses interview evidence of problems in previous jobs), they are likely to respond differently in interviews and render the algorithm worthless.

Below, we address each of these challenges separately at each stage of what we call the AI Life Cycle: Operations – Data Generation – Machine Learning – Decision-Making. We rely on key ideas from Evidence-Based Management (EBMgmt) - a theory-driven causal analysis of "small data" (Barends and Rousseau 2018; Pfeffer and Sutton 2006; Rousseau 2014).

We then suggest how, given these constraints, we might make progress in the application of machine learning tools to HR. Specifically, we focus on the role of causal models in machine learning (Pearl 2009, 2018). We also suggest that randomization can be useful as a decision process, given its perceived fairness and the difficulty that analytics may otherwise have in making fair and valid decisions (Denrell, Fang, and Liu 2015; Liu and Denrell 2018). We base our arguments on knowledge of contemporary practice as well as on interactions with practitioners, and in particular, a 2018 workshop that brought data science faculty together with the heads of the workforce analytics function from 20 major US corporations.

**The AI Life Cycle**

Figure 1 depicts a conventional AI Life Cycle: Operations, Data Generation, Machine Learning, and Decision-Making.

FIGURE 1 HERE

"**Operations**" constitute the phenomenon of interest in its entirety, such as how an organization hires employees. One of the reasons for the interest in applying data science tools to human resources is because HR involves so many operations and so much money is involved in them. In the US economy as a whole, roughly 65 percent of all spending is on labor. In service industries, the figure is much higher. Here are the most common operations in human resources with corresponding prediction tasks for workforce analytics:

| **HR operation** | **Prediction task** |
|---|---|
| Recruiting – identifying possible candidates and persuading them to apply | Are we securing good candidates? |
| **Selection** – choosing which candidate should receive job offers | Are we offering jobs to those who will be the best employees? |
| **On-boarding** - bringing an employee into an organization | Which practices cause new hires to become useful faster? |
| **Training** | What interventions make sense for which individuals, and do they improve performance? |

| | |
|---|---|
| **Performance management** – identifying good and bad performance | Do our practices improve job performance? |
| **Advancement** – determining who gets promoted | Can we predict who will perform best in new roles? |
| **Retention** | Can we predict who is likely to leave and manage the level of retention? |
| **Employee benefits** | Can we identify which benefits matter most to employees to know what to give them and what to recommend when there are choices, and what are the effects of those benefits (e.g., do they improve recruiting and retention)? |

Each of these operations involves administrative tasks, each affects the performance of the organization in important ways, and each includes specific offices, job roles, written instructions and guidelines as well as the actual activities and interactions of all parties. These operations produce volumes of data, in the form of texts, recordings, and other artifacts. As operations move to the virtual space, many of these outputs are in the form of "digital exhaust," which is trace data on digital activities (e.g. online job applications, skills assessment) that may be used to build recruiting algorithms.

Human resource information systems, applicant tracking systems, digital exhaust, and other markers are all critical inputs for the "**data generation"** stage. Typically, this input has to be extracted from multiple databases, converted to a common format, and joined together before analysis can take place.

By "**machine learning**" (ML) we refer to a broad set of techniques that can adapt and learn from data to perform better and better at a task, typically prediction. A data scientist chooses a machine learning algorithm, determines the most appropriate metric to assess its accuracy, and trains the algorithm using the training sample. Some of the most commonly used prediction algorithms, such as logistic regression, random forest, and neural networks infer the outcome variable of interest from statistical correlations among observed variables. The accuracy of preliminary models is assessed on the development data until it stabilizes at some acceptable level.

For hiring, for example, we might see which applicant characteristics have been associated with better job performance and use that to select candidates in the future. As another example, "algorithmic management," the practice of using algorithms to guide incentives and other tools for "nudging" platform workers and contractors in the direction of the contractee (Lee et al 2015), is coming to regular employees as well. At present, this is principally the case in making recommendations. IBM, for example, uses algorithms to advise employees on what training make sense for them to take, based on the experiences of similar employees; the vendor Quine uses the career progression of previous employees to make recommendations to client's employees about which career moves make sense for them. Vendors such as Benefitfocus are able to develop customized recommendations for employee benefits, much in the same way that Netflix can recommend content based on consumer preferences or Amazon can recommend products based on purchasing or browsing behavior.

These algorithms differ in some important ways from traditional approaches used in HR. In industrial psychology, the field that historically focused the most attention on hiring, research would test separate explanatory hypotheses about the relationship between individual predictors and job performance. Machine learning, in contrast, generates one algorithm that may make use of many variables, that may not be in the cannon of the theoretical literature associated with the topic. Indeed, one of the attractions of ML is investigating non-traditional factors because the goal is to build a better prediction, not advance the theory of the field in which the researcher is based.

"**Decision-making**," the final stage, deals with the way in which we use insights from the machine learning model in everyday operations. If anything, employers have more discretion now in how they use the insights from these models than they did in the heyday of the great corporations and Personnel Psychology when hiring and other practices were standardized across an entire company. Managers today typically have the option of ignoring evidence about predictions, using it as they see fit, and generating their own data on practices like hiring in the form of interviews.

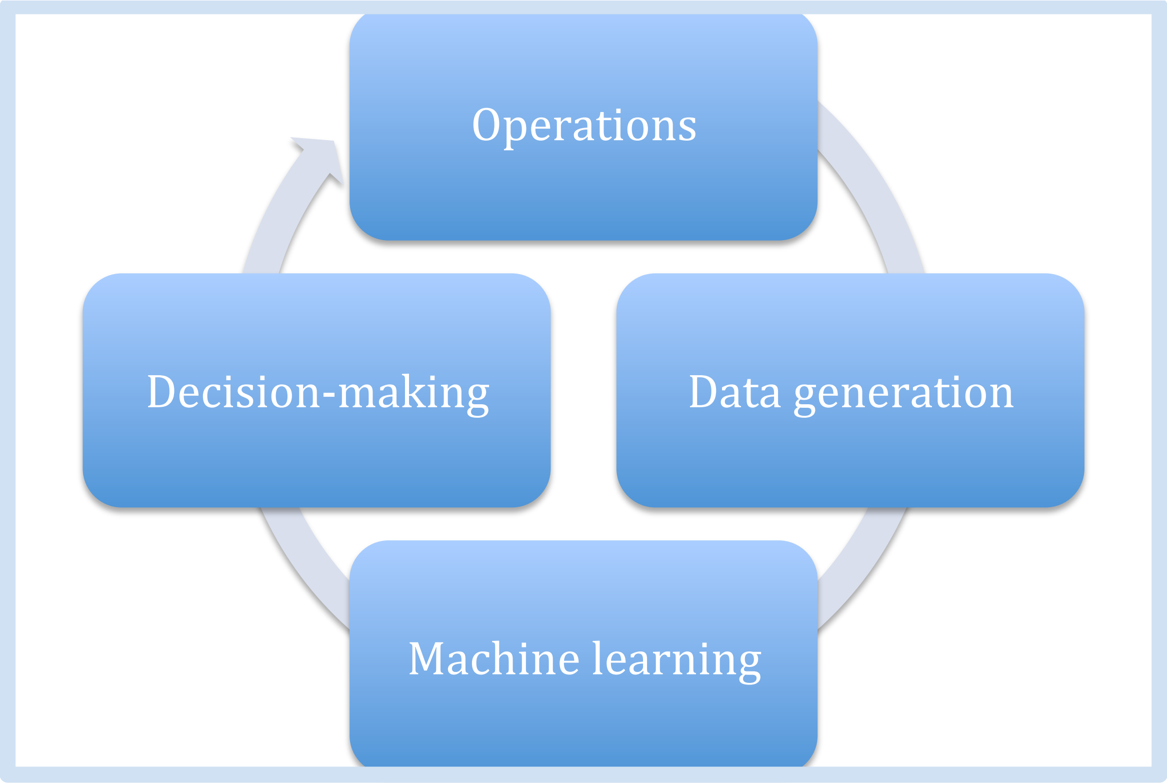


**Figure 1. The life cycle of an AI-supported HR practice**

## Addressing AI Challenges: One Stage at a Time

In this section, we explore in detail the four general challenges to AI outlined in the Introduction: complexity of HR phenomena, small data, ethical and legal constraints, and employee reactions to AI-management. To make these challenges tractable, we discuss them in the context of the particular stages of the AI Life Cycle in which they are most relevant.

### Data Generation stage

The *complexity* inherent in many HR phenomena manifests itself at the Data Generation stage. The most important source of complexity may be the fact that it is not easy to measure what constitutes a "good employee," given that job requirements are broad, monitoring of work outcomes is poor, and biases associated with assessing individual performance are legion. Moreover, complex jobs are interdependent with one another and thus one employee's performance is often inextricable from the performance of the group (Pfeffer and Sutton 2006). Without a clear definition of what it means to be a good employee, a great many HR operations face considerable difficulty in measuring performance.

In terms of the data, not all details of operations leave digital traces, and not all traces left can be extracted and converted to a usable format at a reasonable cost. For example, employers may or may not track the channels through which applicants come to them – from referrals vs. visiting our website vs. job boards, and so forth. Most employers do not retain data on applicants that they screen out. These choices limit the types of analyses that can be performed and the conclusions that can be drawn.

There is no list of "standard" variables that employers choose to gather and to retain through their HR operations. That reduces the extent to which best practices in analytics can be transferred across organizations. Behavioral measures from attitudinal surveys vary considerably in their use across organizations, differences in cost accounting mean that that the detail that employers have on the costs of different operations differs enormously (e.g., are training costs tracked, and if so, how disaggregated are the data?).

The complexity of HR phenomena at the Data Generation Stage is not entirely new. When tackling these challenges, employers can benefit from the key lessons from performance management:

- Do not seek perfect measures of performance, which do not exist anyway. It is better to choose reasonable measures and stick with them than to keep tinkering with systems to find the perfect measure.
- Aggregate information from multiple perspectives and over time. Digital HR tools allow for rapid real-time evaluations among colleagues using mobile devices.
- Use objective measures of performance outcomes based on ex ante determined goals and KPIs whenever possible but complement them with more subjective evaluations to capture less tangible outcomes, such as whether the employee fits into the company's desired culture.
- Integrate HR data with the company's business and financial data to analyze effects on business unit performance.

The complexity of HR phenomena creates another problem in the form of specialized vendors who address only one task. It is very common for an employer to have a system from one vendor to track employee performance scores, from another for applicant tracking software, from a third for compensation and payroll data, and so forth. Aggregating the data for analysis becomes an enormous challenge given that the systems are rarely compatible. It is no surprise that such database challenges were one of the biggest challenges reported by the HR analytics practitioners in our workshop (see Figure 2). In addition to technical barriers, our respondents reported the resistance of other functions to sharing their data with HR Departments.

To illustrate how rudimentary most of the existing database management efforts still are with HR operations, the vast majority of our practitioners reported that the software they most often used to organize and manage their data was Excel. Very few used more purpose-built tools such as Tableau. Software for bridging datasets and “data lakes” that can archive and access to different data sets clearly represent a way forward, but they can be difficult to integrate, can be viewed as confining, and face their own limitations, so they remain under-used in the HR world.

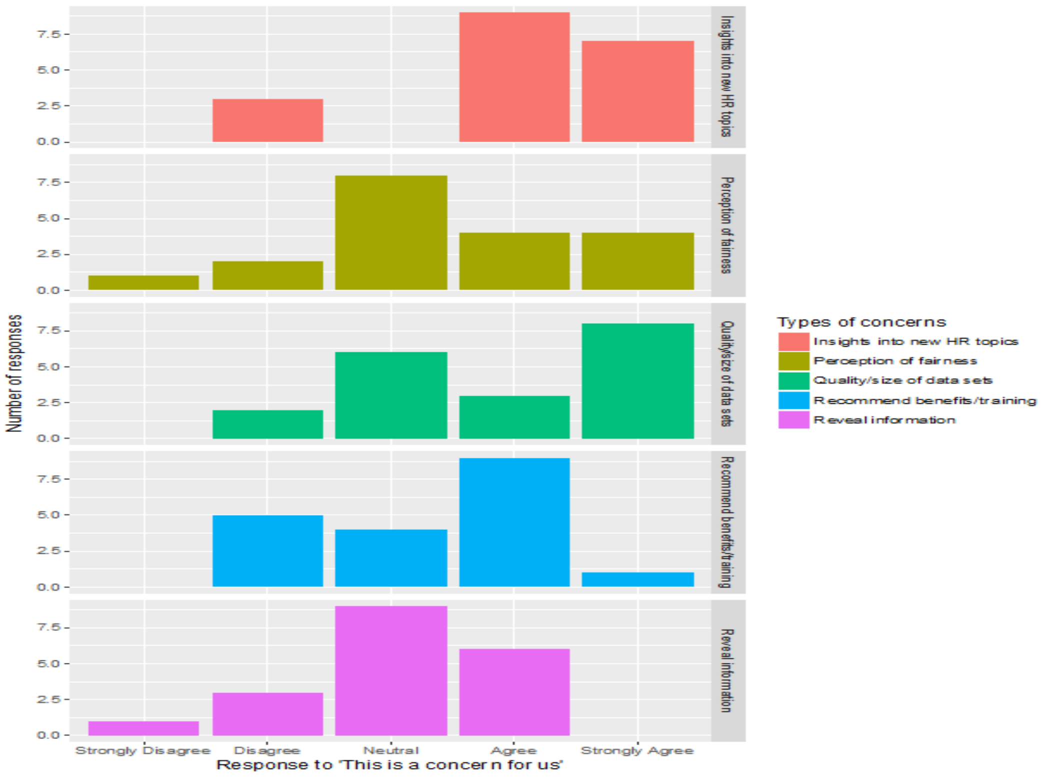
Before launching a major Digital HR project, employers should determine what data are necessary and audit what is available and can be extracted and transferred into a useable format with reasonable costs. For example, if the employer wants to use a machine-learning algorithm in hiring, it needs to collect historical data on job candidates who were not hired. To demonstrate its commitment to digital transformation as well as to benefit from it, companies’ top management has to make data sharing a priority in the short-run and invest in data standardization and platform integration in the long-run.

Given these database concerns, it can be costly to analyze a question in HR for the first time. Data analytics managers, therefore, have to be careful about where to “place bets” in terms of assembling data for analysis, let alone when collecting new data. How should managers decide which HR questions to investigate?

This challenge was the most important concern expressed by our practitioners (see Figure 2). Beyond the obvious criteria of cost is the likelihood of generating useable results. Our practitioners said that in this context, they relied on induction to make the choice: they ask people in HR operations what they have seen and what they think the important relationships are. Some go to senior management and solicit answers to the question of what types of problems prevent the managers from “sleeping at night.” Such experience-driven heuristics are a typical approach under uncertainty. The practitioners also indicated that another factor shaping where they placed their bets is whether anyone was willing to act on results they found.

A more systematic response would include examining the research literature in order to establish what we already know about different research questions, as the evidence-based management has long advocated (Barends and Rousseau 2018). The fact that this approach appears not to be used very often reflects the disconnect between the data science community, which understands analytics but not HR, and the HR community, which understands HR but not analytics. Many leading IT companies, such as Amazon, Google, Facebook, and Microsoft, hire as many PhDs in social sciences as in data sciences to help close this disconnect. In the longer run, AI may be able to parse the research literature itself to identify the questions that can be asked and the models that can be tested given the available data.

FIGURE 2 HERE



*Small Data* is a fundamental concern for human resource analytics. Most employers do not hire many workers, nor do they do enough performance appraisals or collect enough other data points for their current workforce to use machine learning techniques because they do not have that many employees. The machine learning literature has shown that access to larger data has substantial advantages in terms of predictive accuracy (Fortuny, Martens, and Provost 2014).

At the same time, small data are often sufficient for identifying causal relationships, which managers need to understand in order to act on insights from workforce analytics. The management literature has an important advantage over data science in articulating causal relationships, as opposed to prediction from correlations among observed variables. Only recently, some powerful voices in the computer science community articulated the problem of causality as critical for the future of AI in human affairs (Pearl 2018).

The less data we have the more theory we need to identify causal predictors of the outcome of interest. Theory can also come from expert knowledge and managerial experience. AI-management requires that managers put their assumptions on the table, though, and persuade the other stakeholders in their accuracy. The formulation of such assumptions often turns into a contest among stakeholders. This is a place where process formalization that presumes contributions from stakeholders is required.

Where a formal process reveals large disagreements, an alternative might include collecting additional data from randomized experiments in order to test causal assumptions. Google became known for running experiments for all kinds of HR phenomena, from the optimal number of interviews per job candidate to the optimal size of the dinner plate in the cafeteria (Bock 2015). If discussions, experiments, and leadership's persuasion do not lead to a reasonable consensus on the causal model that generates the outcome of interest, AI-analyses are likely to be counterproductive and thus should be avoided until more or better data can be collected.

One attraction of using vendors is their ability to combine data from many employers to generate their algorithms. Such approaches have already been used with standard paper-and-pencil selection tests, or as they are sometimes known now, pre-employment tests, such as those for sales roles. For instance, the well-known company ADP, which handles outsourced payroll operations for thousands of companies, has been able to harness this scale to build predictive models of compensation and churn. Client companies are willing to make their data available for this exercise in return for access to benchmarked comparisons.

The complication for individual employers is knowing to what extent their context is distinct enough that an algorithm built on data from elsewhere will make effective predictions in their own organization. As is discussed further below, such evidence is essential to address legal concerns.

*Employee reactions to data collection efforts.* The fact that employees can bias their responses and the data depending on how they think the data will be used is an important concern for all HR analytic efforts. Because of this, there is concern about finding alternative sources of data that might be viewed as more authentic. A great many employers now make use of social media information on hiring (e.g., looking for evidence of bad behavior, looking for evidence of fit); others use it to assess “flight risk” or retention problems (e.g., identifying updated LinkedIn profiles). Banks have tighter regulations requiring oversight of employees and have long analyzed email data for evidence about embezzlement. They are now using it as well to identify other problems. For example, the appearance of terms like “harassment” in email traffic may well trigger an internal investigation to spot problems in the workplace. (Once employees understand that, of course, they may well change their language either to help generate investigations or prevent them.)

The vendor Vibe, for example, uses natural language processing tools to gauge the tone of comments that employees post on internal chat boards, thereby helping to predict employee flight risk. Applications such as these can face some key challenges when introduced into the workplace. For instance, when employees realize their posts are being used to derive these types of measures, it can influence what and how they choose to post. Then, there are the issues that may arise around whether employees consider such use of the data to infringe upon their privacy.

Several of the companies at our workshop reported that they built models on predicting flight risk and that the best predictors did not come from traditional psychology-based findings but from data sources like social media. Many employers felt that there was an ethical problem with their own use of social media; others felt that data was ok to use but that tracking sentiment on email messages using natural language algorithms was out of bounds; still others thought that any employee-related data was appropriate to use as long as it was anonymized.

Many of these and similar considerations fall under the purview of privacy, which acquires new dimensions in the digital age: data persistence, data repurposing, and data spillovers. (Tucker 2017). Data can persist well beyond the time it was generated and after its use by an employer and employers might use them for purposes unanticipated by the creator, e.g., the words used in an email exchange with a colleague might be used to predict flight risk. It may also inadvertently affect other people, for example, the creators' friends tagged in posts and photos. Here employers have to account for governments' regulations of privacy issues, such as "the right to be forgotten" or the EU's General Data Protection Regulation (GDPR). The former states that business has to satisfy individuals' demands to delete their digital traces after some period of time; the latter is a comprehensive treatment of all the aspects of data privacy in the digital age (www.eugdpr.org).

In terms of technological solutions to the issue of data privacy, computer scientists are actively working on privacy-preserving data analytic methods that rely on the notion of differential privacy in building algorithms. Here data is randomized during the collection process, which leads to "learning nothing about an individual while learning useful information about the population" (Roth 2014: 5). Analysts do not know whose data are used in the analysis and whose data are replaced by noise, but they do know the noise generating procedure and thus can estimate the model anyway. Data on individuals still gets used in applying the algorithm, though: If we find that curse words on social media are negatively related to future job performance, we still have to identify whether a candidate uses curse words on social media to make use of the algorithm.

## **Machine Learning stage**

It may not be surprising that an ML algorithm for predicting which candidates to hire performs better than anything an employer has used before. Indeed, a reasonable complaint is that prior research in human resources has not done much to help employers: the fact that most of the predictors advocated in that research, such as personality and IQ scores, predict so little of job performance (a typical validity coefficient of .30, for example, translates to explaining nine percent of the variance in performance) creates an enormous opportunity for data analytics.

However, finding good data with which to build an algorithm can be challenging. A common approach in the vendor community is to build an algorithm based on the attributes of a client firm's "best performers" and then assess applicants against that algorithm. For instance, the vendor HireVue helps clients conduct video interviews. Part of its offering includes algorithms that analyze a candidate's actions that are captured on video – such as a smile or a particular facial expression – to predict the candidate's future performance at the company. These algorithms are trained on data from other top performers at the client firm. Of course, a potential challenge with using top performers to train algorithms is that it "selects on the dependent variable," by examining only those who are successful. We cannot then identify that which distinguishes best performers from other performers.

Moreover, self-selection can induce spurious relationships among workers' characteristics, which is called the collider effect in data science (Pearl 2018). Many firms inadvertently induce such an effect themselves when they create these "best performer" profiles from an analysis of the determinants of performance of hired workers and apply this profile to job candidates. The analysis usually does not account for the fact that these workers have likely been selected from the pool of candidates on the basis of the same and other determinants that correlate positively with performance. Paradoxically, the same correlations in the sample of hires might be close to zero or even negative, while positive correlations in the same sample might not be observed in the pool of candidates at all. Moreover, the ability of the model to "keep learning" and adapt to new information disappears when the flow of new hires is constrained by the predictions of the current algorithm.

Workers' self-selection into a firm's pool of applicants also creates a biased picture of labor supply for hiring algorithms, especially if the firm has a distinctive brand or identity, positive or negative. As a result, a firm might mistakenly conclude that there is a shortage of some labor and abundance of other. This might prompt an unnecessary loosening or strengthening of hiring criteria.

Several aspects of the modeling process can also be challenging. For instance, there may be more than one measure of "fit" with the data. When we consider the difference between majority populations (e.g., white employees) and minority populations (e.g., African American employees), algorithms that maximize predictive success for the population as a whole may discriminate against predictive success for the minority population. Generating separate algorithms for each, which might lead to better outcomes, conflicts with legal and ethical norms of disparate treatment. As a result, there are fundamental tradeoffs in this domain between accuracy and fairness that must be confronted in any HR ML implementation (Loftus et. al. 2018).

One well-known case of the difficulty in defining fairness in machine learning algorithms is that of the ProPublica investigation into the use of commercial machine learning software by judges in Broward County, Florida to determine whether a person charged with a crime should be released on bail. The tool they used, it was argued, was biased against black defendants. A debate ensued between ProPublica and the developers of the software about the question of whether or not the algorithm was indeed biased. Much of the differences between the two sides ultimately amounted to the fact that there are multiple ways to define fairness in the decision-making process. For instance, one view of fairness might involve excluding the race variable from the algorithm altogether, while another might focus simply on maximizing the likelihood of identifying defendants, regardless of race, who will reoffend. Still another might focus on ensuring that the frequency of misclassifying defendants as reoffenders is not much greater for one race than it is for another. Even when developing the algorithm, therefore, it can be difficult to converge on key issues that have fundamental implications for design.

Many practitioners poorly understand and thus often ignore these issues. Their ignorance might not lead to severe consequences in fields like marketing, where incorrectly targeting ads is not a big deal, but might become very costly in HR where typical decisions, such as hiring, pay, or promotion, have major performance and legal consequences. Having the right populations and samples at the Data Generation stage is only one step towards avoiding biased machine learning.

The other and more comprehensive solution is casual discovery. Computer algorithms of causal discovery are being actively developed, and their interpretation does not require advanced training (Malinsky and Dansk 2017). These algorithms search for causal diagrams that fit the available data. In turn, causally-informed machine learning algorithms allow us to “minimise or eliminate the causal dependence on factors outside an individual’s control, such as their perceived race or where they were born” (Loftus et. al. 2018: 7).

### Decision-Making stage

There are three main challenges when decision makers try to apply the predictions produced by machine learning. The first concerns fairness and legal issues, the second relates to lack of explainability of the algorithm, and the third to the question of how employees will react to algorithmic decisions.

#### *Fairness*

Within the HR context, there are numerous questions related to fairness. One of the most obvious of these is the recognition that any algorithm is likely to be backward looking. The presence of past discrimination in the data used to build a hiring algorithm, for example, is likely to lead to a model that may disproportionately select on white males. Actions using those algorithms risk reproducing the demographic diversity – or lack thereof - that exists in the historical data.

In the HR context, there is a wide-spread belief that evaluations of candidates and employees are shaped heavily by the biases of the evaluator, most commonly as related to demographics. Algorithms can reduce that bias by standardizing the application of criteria to outcomes and by removing information that is irrelevant to performance but that might influence hiring manager decisions, such as the race and sex of candidates. Factors that may be important for the predictive power of the algorithm may nonetheless be seen as inappropriate, such the social status of one’s alma mater. Will that lead us to take out such criteria and lose predictive power?

The possibility of legal challenges especially associated with the adverse impact of employment decisions raises different concerns. Letting supervisors make employment decisions without guidance will likely lead to far more bias than the algorithms generate. But that bias is much harder to hold accountable because it is unsystematic and limited to that workgroup. Algorithms used across the entire organization may have less bias than relying on disparate supervisors, but bias that does result is easier to identify and effects entire classes of individuals, leading to potential class action law suits. Will decision makers find it worthwhile to take that legal risk in order to reduce total bias?

Even if an algorithm addresses concerns about discrimination against protected groups, there are concerns about individual fairness: Suppose the algorithm determined the two best candidates with an 80% and 90% match to the job. Is the 10% difference large or small, accounting for the algorithm's accuracy? The hiring manager might review all available information and make an expert judgment. However, there is enough evidence that such judgments are often just post-rationalization that introduces its own biases and random noise.

One way to mitigate some of these issues is by introducing random variation, which has been an unrecognized but important mechanism in management (Denrell, Fang, and Liu 2015; Liu and Denrell 2018). Research shows that employees perceive random process as fair in determining complex and thus uncertain outcomes (Lind and Van den Bos 2002). Therefore, if the algorithm is sound and both candidates are strong, it makes more sense to accept the complexity of hiring decisions and make a random choice between the two candidates, possibly, with the probabilities of selection proportional to the matching scores. In other words, randomization should be an AI-management tool at the Decision-Making stage as much as at the previous stages of the AI Life Cycle.

Of course, context matters. Many of our participants found it perfectly acceptable to use algorithms to make decisions that essentially reward employees – who to promote, who to hire in the first place. The inevitable use of algorithms to punish employees raises more questions, though: An algorithm that predicts future contributions will most certainly be introduced at some point to make layoff decisions, but how about one that predicts who will steal from the company or commit a crime?

Here we face a dilemma.  The Utilitarian notion of advancing the collective good might well argue for using predictive algorithms to weed out problems and costly employees.  The Kantian deontological position, on the other hand, suggests that individuals should be judged based on their own actions and would find that position highly objectionable.

Related to the concept of fairness is the acceptability that engagement in decision making generates for most individuals. Research increasingly shows that algorithms perform better than human judgment when used to predict repetitive outcomes, such as reading x-rays but also predicting outcomes about employees or job candidates (Cowgill 2018). But if algorithms take over hiring and supervisors play no role in the process, will they be as committed to the new hires as if they had made the hiring decisions?

*Explainability*

With machine learning models, the ability to explain what factors the algorithm is using to make predictions is generally much more difficult than it is with traditional statistical models. A simple seniority decision rule – more senior workers get preference over less senior ones – is easy to understand and feels objective even if we do not like its implications. A machine learning algorithm based on a weighted combination of 10 performance-related factors may be just as objective as a traditional statistical model but is much more difficult to understand, especially when employees make inevitable comparisons with each other. Algorithms get more accurate the more complicated they are, but they also become more difficult to understand and explain.

A well-known example of the importance of explainability to users comes from the Oncology application of IBM Watson. This application met considerable resistance from oncologists because it was difficult to understand how the system was arriving at its decisions. When the application disagreed with the doctor's assessment, this lack of transparency made it difficult for medical experts to accept and act upon the recommendations that the system produced. Especially in "high stakes" contexts, such as those that affect people's lives—or their careers--explainability is likely to become imperative for the successful use of machine learning technologies.

*Employee reactions to algorithmic decisions*

Changes in formal decision-making unavoidably affect employees' behavior. In this regard, we can learn a great deal from Scientific Management's efforts to develop optimal decision rules. Employment practices and decisions about work organization were based on a priori engineering principles and human experiments. Although they may have been much more efficient than previous practices, they were bitterly resented by workers, leading to a generation of strife and conflict between workers and management. From the perspective of front-line workers and their supervisors, the situation may have looked very similar to the AI model we outline here: decisions would been handed down from another department in the organization, the justification for them would be that they were the most efficient that science could provide, and trying to alter them would simply be a mistake.

To illustrate a likely consequence, it is widely believed that the relationship with one's supervisor is crucial to the performance of their subordinates and that the quality of that relationship depends on social exchange: "I as supervisor look after you, and you as subordinate perform your job well." Even when employees have little commitment to their employer as an organization, they may feel commitment to their supervisor. How is this exchange affected when decisions that had been made by the supervisor are now made by or even largely informed by an algorithm rather than a supervisor?

In a workplace context, if my supervisor assigns me to work another weekend this month, something I very much do not want to do, I might do it without complaint if I think my supervisor has been fair to me. I might even empathize with the bind my supervisor is in when having to fill the weekend shift. If not, I might well go complain to her and expect some better treatment in the future. When my work schedule is generated by software, on the other hand, I have no good will built up with that program, and I cannot empathize with it. Nor can I complain to it, and I may well feel that I will not catch a break in scheduling in the future. We know, for example, that people respond very differently to decisions that are made by algorithms than decisions made by people (Dietvorst, Simmons, and Massey 2016). If there is good news to give me, such as a bonus, it builds a relationship with my supervisor if she appears to have at least been involved in the decision, something that does not happen if that decision is generated by an algorithm.

Yet, there may be occasions where decisions are easier to accept when made by an algorithm than when made by a human, especially when those decisions have negative consequences for us. Uber riders, for example, respond negatively to surge pricing increases when they perceive that they are set by a human (trying to exploit them) as opposed to by an algorithm. Experimental evidence suggests that willingness to accept and use algorithms depends in part on how they update to deal with mistakes (Dietvorst et al forthcoming).

**Discussion and Conclusions**

While general-purpose AI is still a long shot in any domain of human activity, the speed of progress towards specialized AI systems in health care, automobile industry, social media, advertising and marketing is  considerable.  Far less progress has been made in issues around the management of employees.   We identify four reasons why: complexity of HR phenomena, small data from HR operations, fairness and legal constraints, and employee reactions to AI-management.

Causal reasoning is the first principle relevant to addressing these challenges across  the stages of the AI Life Cycle as it  makes practical takeaways from small data possible and  algorithm-supported decision-making  fairer and explainable. Of course, these benefits do not come without costs: Companies have to accept algorithms' lower predictive power and seek some consensus about causal assumptions. The former is the main reason why the data science community at large is quite skeptical about causally reasoning AI systems.

Randomization is a second principle that can help with establishing causality and may also be a decision rule for when faced with low predictive power from algorithms, enhancing the perception of fairness among employees.

Finally, formalizing processes is necessary to ensure that the assumptions built into algorithms are reasonable and that the outcomes are understandable.  In the process, formalization can be enabling rather than coercive (Adler and Borys 1996).

To what extent the changes we suggest require a restructuring of the HR function is an important question. Certainly HR leaders need to understand and facilitate the Data Generation and Machine Learning stages of the AI Life Cycle. The integration of HR data with business and financial data should allow an HR Department to quantify in monetary terms its contribution to the company's bottom-line.

Line managers will have to refresh their skill set as well. For them, AI should stand for "augmented intelligence," an informed use of workforce analytics' insights in decision-making. The literature on evidence-based management proposes a Bayesian approach to systematically updating managerial beliefs with new information (Barends and Rousseau 2018). We consider it as a helpful departure point for AI-management as well.

The tension between the logic of efficiency and of appropriateness affects most organizational action (March and Simon 1993). In the case of HR, the drive for efficiency and concerns about fairness do not always align. We hope that the conceptual and practical insights in this paper will move AI-management in HR forward on both counts, those of efficiency and appropriateness.

**Bibliography**

Acktar, Reese, Dave Winsborough , Uri Ort , Abigail Johnson , Tomas Chamorro-Premuzic. 2018. Detercting the Dark Side of Personality Using Social Media. Personality and Individual Differences, 132:90-97.

Adler, Paul and Bryan Boris. 1996. "Two Types of Bureaucracy: Enabling and Coercive." *Administrative Science Quarterly* 41: 61-89.

Barends, Eric and Denise M. Rousseau. 2018. *Evidence-Based Management: How to Use Evidence to Make Better Organizational Decisions.* Kogan Page.

Bock, Laslo. 2015. *Work Rules! Insights from Inside Google That Will Transform How You Live and Lead.* Hachette Book Group.

Cappelli, Peter. 2017. "There's No Such Thing as Big Data in HR." *Harvard Business Review*. June.

Cappelli, Peter and AnnaTavis. 2017. The Performance Management Revolution. *Harvard Business Review,* November.

Carrillo-Tudela, C., Hobijin, B., Perkowski, P., and Visschers, L. 2015.

Denrell, Jerker, Christina Fang, Chengwei Liu. 2015. "Change Explanations in Management Science." *Organization Science* 26(3): 923-940.

IBM 2018.

Cowgill, Bo (2017) The Labor Market Effects of Hiring through Machine Learning Working Paperowgill, Bo. 2018. "Bias and Productivity in Humans and Algorithms. Theory and Evidence from Résumé Screening." Working paper.

Dietvorst, Berkeley, Simmons, Joseph P. ,and Massey, Cade, Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err (July 6, 2014). Forthcoming in *Journal of Experimental Psychology: General.*

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155-1170.

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). "Predictive modeling with big data: is bigger really better?" *Big Data*, *1*(4), 215-226.

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. 2015. *Working with machines: The impact of algorithmic, data-driven management on human workers*. Proceedings of the 33rd Annual ACM SIGCHI Conference: 1603-1612. Begole, B., Kim, J., Inkpen, K & Wood, W (Eds.), New York, NY: ACM Press.

Lind, E. Allan and Kees Van den Bos. 2002. "When Fairness Works: Toward a General Theory of Uncertainty Management." *Research in Organizational Behavior* 24: 181-223.

LinkedIn. 2018. The Rise of HR Analytics.

Liu, Chengwei and Jerker Denrell. 2018. "Performance Persistence Through the Lens of Chance Models: When Strong Effects of Regression to the Mean Lead to Non-Monotonic Performance Associations." Working paper.

Loftus, Joshua R., Chris Russel, Matt J. Kusner, and Ricardo Silva. "Causal Reasoning for Algorithmic Fareness." **arXiv:1805.05859**

Malinsky, Daniel and David Danks. 2017. "Causal Discovery Algorithms: a Practical Guide." *Philosophy Compass* https://doi.org/10.1111/phc3.12470.

March, James and Herbert Simon. 1993. *Organizations.* Oxford: Blackwell.

Pearl, Judea. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pfeffer, Jeffrey and Robert I. Sutton. 2006. *Hard Facts, Dangerous Half-Truths and Total Nonsesne: Profiting from Evidence-Based Management.* Harvard Business Review Press.

Rousseau, Denise (Editor). 2014. *The Oxford Handbook of Evidence-Based Management.* Oxford University Press.

Srivastava, Sameer and Amir Goldberg. 2017. “Language as a Window into Culture.” *California Management Review* 60(1): 56-69.

Tucker, Catherine. 2017. “Privacy, Algorithms, and Artificial Intelligence.” In *The Economics of Artificial Intelligence: An Agenda.* Edited by Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb. University of Chicago Press. Forthcoming. http://www.nber.org/chapters/c14011.